Автоматическая классификация текстов на основе их структурных признаков. Какую информацию о тексте отражает структура?

© Пустыльникова О. Мелер А.

Университет Билефельд, Германия

Olga.Pustylnikov@uni-bielefeld.de
Aleksander.Mehler@uni-bielefeld.de

Аннотация

This paper presents an algorithm to automatically classify text documents into thematic fields. The algorithm operates only on text structure disregarding any content information. We present an evaluation of the approach using the SUSANNE corpus [1] of written English and LUCY, a corpus of adult, child and young adult writing [1]. We show that using only a small number of features it is possible to achieve good classification results.

1. Введение

Одной из задач информационного поиска является разработка наиболее эффективных методов, использующих наименьшее количество информации о данных. Наиболее эффективными в этой области являются методы, основанные на подборе характеристик «*bag of features*» [2], позволяющих сравнивать и классифицировать тексты. В большинстве случаев характеристиками являются частоты отдельных слов, которые помогают выявить тематические категории, к которым принадлежат данные тексты. Чтобы получить хорошие результаты, нужно иметь огромное количество таких характеристик. С одной стороны этот метод дает хорошие результаты, с

другой стороны хотелось бы получить те же результаты, используя наименьшее количество признаков.

Частотные характеристики слов позволяют проводить тематическую классификацию текстов. А что, если категории определены не тематически, а, например, по возрасту людей, написавших тексты. При этом, темы текстов могут быть совершенно разными, а стиль написания может быть схож. Авторы [13,14], работавшие над стилем, заметили, что помимо частот слов существуют и другие уровни информации о тексте, которые можно использовать для классификации.

Такими уровнями могут служить, например, оформление текста, или его структура. Под структурой, мы понимаем как логическую (разделение на главы, параграфы, предложения, заголовки и т.п.), так и синтаксическую структуру текстов. Методы «bag of words», сокращающие текст до набора слов, оставляют эту информацию без внимания.

В данной работе мы подходим к информационному поиску с другой точки зрения и пытаемся классифицировать тексты, основываясь исключительно на их структуре (распределение частоты фраз, предложений и т.п., см. ниже). Под структурой мы подразумеваем расположение фраз, предложений, а также заголовков, абзацев и проч. структурных элементов и их статистическое распределение. Ведь чтению и пониманию текста способствует не только понимание отдельных слов, но и построение предложения, которое часто помогает определить или угадать данный смысл до прочтения всего предложения.

Согласно гипотезе Бибера [3], формы речи (и текста) используются в соответствии с жанром, который «диктует» выбор той или иной формы. Предыдущие исследования, использовавшие SUSANNE, а также газетный корпус немецкого языка, содержащий 95 рубрик [4,5,6], подтверждают эту гипотезу.

В данной работе, тексты принадлежат к нескольким категориям, относящимся как к жанрам языка (художественный, научный язык и т.п.), так и к другим категориям (типа подростковый, детский язык). Мы хотим узнать, удастся ли разделить тексты на эти группы используя их структуру. Результаты исследований должны помочь ответить на вопрос, что именно отражает структура текста, и какие ее возможности в отношении к разным по качеству категориям.

Мы тестируем классификатор четырьмя способами, используя кластерный анализ (unsup), метод опорных векторов (svm) [7], генетический алгоритм по подбору лучших признаков (genetic) и случайный кластерный анализ (random) в качестве проверки. Глава 2 описывает используемый квантитативно-структурный метод (Quantitative Structure Analysis), позволяющий выделить и использовать структурные признаки. В треть-

ей главе представлены данные и их описание. В главе 4 даны результаты. В главах 5 и 6 мы подведем итоги статьи.

2. Метод

2.1 Квантитативно-структурный анализ (QSA)

Главная идея квантитативно-структурного метода состоит в том, что структура текста (будь то заголовки, абзацы, предложения, фразы и т.п.) коррелирует с его функцией, содержанием и жанром [3, 12]. Методы, использующие исключительно частоту смысловой лексики, используют только часть информации, заложенной в текстах. Метод QSA основан на векторном представлении текстов, компонентами которых служат частоты элементов структуры. Этот метод основан на работах Тулдавы [8], который использовал кластерный анализ для классификации текстов на базе простых квантитативных признаков. С другой стороны в области синергетической лингвистики Кёлер [9] предложил ряд квантитативных параметров, применяемых к синтаксическим структурам. Эти характеристики используются QSA для характеристики синтаксической и логической (logical document structure[1]) структуры текстов.

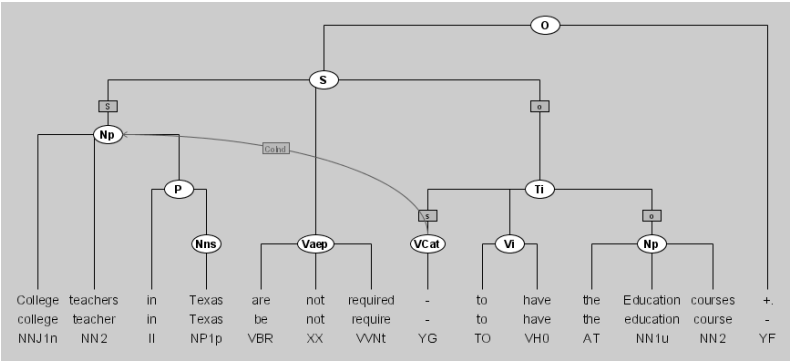

Рис. 1.

[1] Под логической структурой текста подразумевается его Document Object Model (DOM) или текст, представленный как дерево из глав, параграфов, предложений и т.п.

2.2 Квантитативно-структурный профиль текста

Структурные признаки, разработанные Кёлером, применяются к синтаксическим деревьям и изначально подразумевают наличие синтаксической разметки текстов. Однако, в случае немецкого газетного корпуса использовалось дерево логической структуры текстов, что не повлияло на результат классификации. В [6] работе показано, что QSA дает лучшие результаты, чем другой комплексный метод, основанный на SVM tree kernels.

В данной работе используются два корпуса с синтаксической разметкой (см. Гл. 3), в которых каждое предложение представлено в виде дерева, разделенного на фразы. Пример такого дерева изображен на рис. 1.

Аннотация корпусов содержит 18 различных структурных типов (например: номинальная фраза (Np), предложение (S) и др., см. рис. 1). Для каждого такого типа $T = \{T1, \ldots, Tm\}$ можно рассмотреть, например, его частоту, а также:

- сложность (**complexity**): как количество дочерних элементов данной фразы. На рис. 1, например, комплексность предложения (S) равна трем.
- длину (**length**): как количество «листьев», т.е. самых нижних элементов (слов), включенных в данную фразу. Например, на рис. 1 длина (S) равна 13, т.е. количество слов, входящих в предложение.
- высоту (**depth**): как количество шагов от данной фразы, до вершины дерева (см. рис.1, от (P) до (S) длина равна 2)
- и другие возможные величины, применяемые к деревьям.

Итак, для каждого структурного типа T вычисляется ряд характеристик (длина, высота и др.) $F = \{F1, \ldots, Fn\}$ встречающихся в одном из текстов $x \in C$. Каждая из характеристик представлена в виде вектора значений, т.е., для структурного типа (S) получается вектор длин, вектор $v \in V$ высоты и т.д.

Чтобы сравнивать тексты, необходимо объединить полученные значения в одно число, представляющее длину, комплексность и т.п. структурного типа Ti в тексте xj. Для этого ко всем векторам v содержащим значения F применяется агрегационная функция $o \in O$. Такой функцией может служить *среднее арифметическое значение*, *энтропия* и другие. В данной работе были использованы три функции: среднее арифметическое значение, стандартное отклонение и энтропия.

Таким образом, получается четерехшаговая процедура, в результате которой, для каждого текста составляется квантитативно-структурный профиль (quantitative structure profile QSP), который может быть использован для классификации. Обобщим еще раз основные шаги процедуры:

1) выбираем структурные признаки *T*;
2) обозначаем квантитативные характеристики *F*;
3) заполняем векторы *V* наблюдениями каждого квантитативного признака *F* для структурного типа *T*i в тексте *x*j;
4) применяем к каждому вектору из *V* функции из *O*, чтобы получить квантитативно-структурный профиль текста, выражаемый вектором Q(*x*j) для каждого текста *x*j.

2.3 Классификация

Когда все тексты представлены в виде векторов Q*x*j, их можно классифицировать. В данном исследовании мы различаем 4 метода классификации:

- кластерный анализ
- support vector machines
- genetic feature selection algorithm
- случайный кластерный анализ (baseline)

Таблица 1

SUSANNE		
Категория	**Количество документов**	**Количество слов/документ**
Belles lettres, memoirs (G)	16	2000
presse reportage (A)	16	2000
scientific writing (J)	16	2000
adventure, fiction (N)	16	2000

3. Данные

Тексты корпуса SUSANNE состоят из четырех гомогенных категорий соответствующих *мемуарному*, *газетному*, *научному и приключенческому* жанру английского языка.

В корпусе LUCY собраны тексты на английском языке написанные взрослыми, подростками и детьми. Взрослая проза в свою очередь разделена на две подкатегории: информационные и художественные тексты. Детская проза включает в себя тексты из четырех подгрупп: сочинения, написанные детьми 9, 10, 11 и 12 лет.

Таблица 2

LUCY			
Главная категория #3	**Субкатегория #7**	**Количество документов**	**Количество слов/документ**
polished	#2	41	101,000
	informative	34	84,000
	imaginative	7	17,000
young adult writing	#1	48	33,000
child writing	#4	150	30,000
	12-year-olds	37	8000
	11-year-olds	36	7000
	10-year-olds	29	6000
	9-year-olds	49	9000

4. Результаты

Таблица 3

Результаты: SUSANNE		
Количество категорий	**метод**	**F-Measure**
4	clustering	0.84591
4	svm	0.85946
	genetic	0.89328
4	random	0.35313

Таблица 4

Результаты: LUCY		
Количество категорий	**метод**	**F-Measure**
3	unsup	0.75348
3	svm	0.81305
3	genetic	0.784
3	random	0.49766
7	unsup	0.41906
7	svm	0.1915
7	genetic	0.4290
7	random	0.23415

Таблица 5

Категория	**F-Measure**
1. young adult writing	1.0
2. 10-year-olds	1.0
3. informative	0.95
4. imaginative	0.88
5. 9-year-olds	0.67
6. 11-year-olds	0.39
7. 12-year-olds	0.33

5. Дискуссия

Результаты таблицы 3 показывают, что возможно классифицировать жанры, с помощью структуры текстов. *F-Measure* 0.8 для SUSANNE у всех трех методов превышает случайную кластеризацию.

В первом случае с LUCY (3 категории) классификация удается (~0.7) не смотря на то, что мы имеем дело не с жанрами, или тематическими категориями, а с текстами, подобранными по возрастным критериям.

Классификация их подкатегорий, однако, дает слабые результаты (0.4). Причиной этого, предположительно, является плохая разграничимость этих категорий. Дальнейшие результаты показали, что, если классифицировать детскую прозу на четыре подгруппы, результат F-Measure не превышает 0.4. Из этого следует, что структурно эти подгруппы различаются слабо.

6. Заключение

Структурно-квантитативный метод позволяет классифицировать тексты, принадлежащие к различным жанровым и другим смысловым категориям. С его помощью возможно установить степень определенности некоторых категорий. В случае детской прозы детей 9, 10, 11 и 12 лет, различия в синтаксической структуре оказываются минимальными. По крайней мере они не достаточны, чтобы различать с их помощью эти четыре группы.

Чтобы выяснить, какие категории наиболее гомогенны и отличны друг от друга, мы провели итеративное упорядочение категорий по лучшему F-Measure [4]. В этой процедуре, начиная с одной базовой категории, пошагово добавляется та категория, которая при добавлении дает лучший результат классификации. Результаты упорядочения показывают, что подростковая проза хорошо отделяется от детской и от взрослой. Однако детские подгруппы 9, 11 и 12 лет с трудом удается разделить, что и портит результат классификации.

Таким образом, нам удалось выяснить, что структурный метод применяем не только к хорошо разделенным жанрам или тематическим категориям, но и к таким разнородным группам какие представлены в LUCY. Условием является отличимость таких групп на структурном уровне. Слабые результаты с детскими подгруппами свидетельствуют о том, что эти подгруппы в действительности мало отличаются друг от друга, либо эти различия настолько неощутимы, что структурные характеристики их не охватывает. Возможно, эти слабые различия можно выявить более высокоуровневыми методами, например, методом представления текстов в виде графов [10, 11].

Итак, структура кроет в себе дополнительную информацию, относящуюся к стилю и жанру текстов, и не только к их тематике. В данной работе мы продемонстрировали эффективность структурного метода на жанровых и на возрастных категориях. Однако метод может быть также применен к веб-жанрам (см. http://ariadne.coli.uni-bielefeld.de/indogram/).

Литература

[1] Sampson, G.R. 1995. English for the Computer: the SUSANNE Corpus and analytic scheme. Clarendon Press (Oxford).

[2] Baeza-Yates, R. and B. Ribeiro-Neto (1999). Modern Information Retrieval. Reading, Massachusetts: Addison-Wesley.

[3] Biber, D. (1995). Dimensions of Register Variation. A Cross-Linguistic Comparison. Cambridge University Press.

[4] O. Pustylnikov and A. Mehler. Structural differentiae of text types. A quantitative model. In Proceedings of the GfKl, 2007.
[5] O. Pustylnikov. Guessing Text Type by Structure. In Proceedings of the ESSLLI Student Session 2007, 2007.
[6] A. Mehler, P. Geibel, and O. Pustylnikov. Structural clasfiers of text types: Towards a novel model of text representation. LDV Forum, 22(2):51{66, 2007.
[7] T. Joachims. Learning to classify text using support vector machines. Kluwer, Boston/Dordrecht/London, 2002.
[8] J. Tuldava. Probleme und Methoden der quantitative-systemischen Lixikologie. Wiessenschaftlicher Verlag, Trier, 1998.
[9] R. Köhler. Syntactic Structures: Properties and Interrelations. Journal of quantitative Linguistics, 1999.
[10] O. Pustylnikov and A. Mehler. Discovering Language Typology by means of quantitative network analysis. To appear: 2008.
[11] Ferrer i Cancho, R., Mehler, A., Pustylnikov, O., and Dıaz Guilera, A. Correlations in the organization of large-scale syntactic dependency networks. In *TextGraphs-2: 2007.*
[12] K. Brinker. Linguistische Textananlyse. Erich Schmidt Verlag, 1985.
[13] Stamatatos, E., G. Kokkinakis und N. Fakotakis. Automatic text categorization in terms of genre and author. Computational Linguistics, 2000.
[14] J. Karlgren. Non-Topical Factors in Information Access. In WebNet (1), 1999.